DOCUMENT RESUME

ED 371 016 TM 021 647

AUTHOR Stone, Gregory Ethan; Lunz, Mary E.
TITLE Item Calibration Considerations: A Comparison of Item Calibrations on Written and Computerized Adaptive Examinations.
PUB DATE Apr 94
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
AVAILABLE FROM Mary E. Lunz, American Society of Clinical Psychologists, 2100 West Harrison Street, Chicago, IL 60612.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Comparative Analysis; *Computer Assisted Testing; Diagrams; Difficulty Level; Illustrations; *Item Response Theory; Medical Technologists; Photographs; Test Format; Testing; *Test Items; *Test Reliability
IDENTIFIERS Calibration; Comparability; *Item Calibration; Item Stability; Rasch Model

ABSTRACT

This paper explores the comparability of item calibrations for three types of items: (1) text only; (2) text with photographs; and (3) text plus graphics when items are presented on written tests and computerized adaptive tests. Data are from five different medical technology certification examinations administered nationwide in 1993. The Rasch model was used to calibrate items for the two test formats. Item calibrations obtained from each administrative mode were then compared. No significant differences were found between text only item calibrations obtained from the written tests and the computerized adaptive test. While some items with photographic or figure accompaniment showed slightly different item calibrations between the administrative modes, nonstatistical explanations explain most of the minor differences discovered. The results of this investigation confirm that Rasch item calibrations from written tests are appropriate for use on computerized adaptive tests. Included are four tables and six figures. (Contains 7 references.) (Author/SLD)

Item Calibration Considerations: A Comparison of Item Calibrations on Written and Computerized Adaptive Examinations

Gregory Ethan Stone

HAB Associates, Incorporated

Mary E. Lunz

American Society of Clinical Pathologists

Copies available from
Mary E. Lunz
American Society of Clinical Pathologists
2100 West Harrison Street
Chicago, Illinois 60612

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, April, 1994.

Item Calibration Considerations: A Comparison of Item Calibrations on Written and Computerized Adaptive Examinations

Abstract

This paper explores the comparability of item calibrations for three types of items, 1) text only, 2) text with photographs, 3) text plus graphics when items are presented on written texts and computerized adaptive tests. Data are from five different medical technology certification examinations administered nationwide in 1993. The Rasch model was used to calibrate items for the two test formats. Item calibrations obtained from each administrative mode were then compared. No significant differences were found between text only MCQ item calibrations obtained from the written tests and the computerized adaptive test. While some items with photographic or figure accompaniment showed slightly different item calibrations between the administrative modes, non-statistical explanations explain most of the minor differences discovered. The results of this investigation confirm that Rasch item calibrations from written tests are appropriate for use on computerized adaptive tests.

Item Calibration Considerations: A Comparison of Item Calibrations on Written and Computerized Adaptive Examinations

Green (1988) suggests that item equivalency between CAT and written administrations is really a problem of scaling rather than equating. Even when there is a shift in item calibrations due to mode of administration, the scales of the tests are equivalent in construct. If results are reproducible in either mode, we can assert that the modes of administration produce comparable results.

The equivalency of item calibrations may also be addressed in terms of stability. In their work on item calibration stability across modes of administration, Bergstrom and Lunz (In press) recalibrated items using data from a computerized adaptive test and compared those recalibrations to the calibrations obtained for the same items from traditional written test administration. Ninety-eight percent of the item calibrations remained stable across administration modes when the shift in scale (standard deviation) between CAT and written was accounted for.

One of the specifications of the Rasch model (Rasch, 1960; Wright & Stone, 1979) is that measurement is independent of the specific set of items used to measure ability and thus the same conclusions should be made about a person's ability regardless of the subset of items taken. This is extremely important in CAT because examinees take different subsets of items, all of which are calibrated to the same scale. Thus the stability of item

calibrations is critical for consistent interpretation of examinee performance.

While text-only MCQ item calibrations obtained from computerized adaptive and written tests have been shown to be highly stable (Bergstrom & Lunz, In Press) little has been done to investigate the performance comparability of MCQ items with accompanying graphical representations presented on the computer screen along with the item. Do these MCQ items with screen graphics function comparably in computerized and written formats?

This paper explores the comparability of item calibrations when presented on screen and in written format of three different types of items: 1) text only, 2) text plus photographs, and 3) text plus non-photographic figures or charts. The compared item calibrations are from two different administrative modes, written and computerized adaptive. Three research questions are addressed. First, are item calibrations for text only MCQ items (which include no visual/graphical presentations) comparable when calculated from written and computerized adaptive response data. Second, are item calibrations for MCQ items with accompanying visual photographs comparable on written tests (with photographs printed in a book) and on computerized adaptive tests (with photographs shown on the screen). Third, are MCQ item calibrations with accompanying graphics (charts, tables, non-photographic figures) from written test data (with the figure printed in a book) comparable to item calibrations from computerized adaptive test data (when the figure is shown on the screen)?

Methods

Data

Data are from three different certification examinations that were administered in 1993. Each examination was administered in both the written and computerized adaptive format. Item calibrations for the written tests were derived from large groups of students who took each item in the written format (Test A = 1052; Test B = 731; Test C = 641). Because of the nature of computer adaptive testing (examinees may not all see the same items) smaller numbers of examinees saw each item in the CAT format. Item calibrations from the CAT exam were derived from varying numbers of examinee responses, minimum of 17-20, maximum of 35.

The exams were high stakes, so examinees were highly motivated to be successful. Three different examinations were analyzed to achieve more generalizable results and avoid the possibility of identifying test specific patterns. This also increased the number of items which presented visual material. The total number of items compared was 54 (20 text only MCQ, 34 MCQ with accompanying figures or photographs).

Design

The Rasch model was used to calibrate items for the two test formats, written and computerized adaptive, after the testing was completed. Mean and standard deviation differences were accounted for by a linear transformation that placed the written item calibrations on the same scale as the calibrations from the CAT:

$$y=a+b(x) \quad (1)$$

where a = intercept, b = slope, x = original measure and y = transformed measure.

This corrected for any "differences of scale" that may have existed (see Tables 1, 2 & 3 for exact formulae).

Item calibrations obtained from each administrative mode were compared using standardized differences calculated as:

$$z=\frac{(d_1-d_2)}{\sqrt{(s^2{}_1+s^2{}_2)}} \quad (2)$$

where d = item calibration (difficulty) and s = standard error.

Written and CAT calibrations for each group of items were then plotted using 95% quality control lines to identify the items with significantly different item calibrations. Where differences were found, the items were examined.

In addition, for purposes of qualitative discussion, text items with non-photographic figures or charts were categorized into three levels of complexity (see figure 1 for specific examples).

Insert Figure 1 about here

Simple figures were those X-Y graphs without number markings and figures with little or no labelling. Average figures were typical X-Y graphs and figures with small or complex labelling. Complex figures included large tables of numbers, complex graphs and very fine print. Our hypothesis was that if calibration differences

appeared, they would most likely involve items with complex figures.

Results

Most 'text only' item calibrations obtained from the written and computerized adaptive administrative formats were comparable, and fell within the 95% confidence band. Two sets of 10, Text only MCQ items, were drawn for each comparison. Item calibrations for each group are presented in Tables 1 and 2 and are compared in Figures 2 and 3 respectively.

Insert Tables 1 & 2 and Figures 2 & 3 about here

When scales were adjusted for variance, item calibrations for most items on both modes fell within the standard error of measurement, as expected.

Insert Table 3 and Figure 4 about here

Table 3 presents calibrations for 13 MCQ items with accompanying figures or photographs drawn from three different tests (a,b,c). Figure 4 shows that of these items with photographs or figures, only 4 showed different item calibrations between the written and computerized adaptive modes. The content and format of these items were reviewed.

Items A and B appeared more difficult on the written test than in the on-screen computerized adaptive mode. Item A was accompanied

by a photograph which was larger on the screen than in print. The enlargement made the image clearer, thus making the item easier. It could be argued that because of the qualitative difference in the clarity of the photograph presented, Item A should instead be considered as a different item in each of the administrative modes. Item B was accompanied by a photograph that was of relatively poor quality in both written and computer presentations. However, when content experts evaluated the item they concluded that it could be answered effectively without the aid of the graph.

Items C and D appeared slightly more difficult when presented on the screen in the computerized adaptive mode than on the written test. For item C, no identifiable reason for a difference could be found. Item D was accompanied by an X-Y plot of average complexity. Perhaps figures, tables and other representations that include numbers are looked at on-screen in a different way than in print. We hypothesized that reading numeric tables and charts on screen may be more difficult that in print.

To test this hypothesis, additional items with charts, tables and other non-photographic accompaniments were selected from across three tests (see Figure 5). Table 4 presents calibrations for MCQ items with accompanying charts, tables or non photographic material, across these three tests (a,b,c).

Insert Table 4 and Figure 5 about here

Written and CAT calibrations for the majority of the MCQ items

accompanied by non-photographic figures or charts had comparable item calibrations. In general, the items appeared equally difficult in the written and computerized adaptive mode. The level of each figure (complex, average and simple) is noted in Figure 5. It is clear that the simple items (labeled S) have comparable calibrations between the two administrative modes. Average items (labeled M) and complex items are generally equivalent. Only one average and one high complexity items had significantly different calibrations across modes of administration.

Two items were identified as calibrating differently in the CAT and written modes. These items (see figure 5) are located to the right of the 95% confidence band, suggesting that they were more difficult on CAT than on the written test. One item (labelled M) is described as average and the other (labelled H) as complex. It was determined that the average item (M) was answerable without reference to the figure. The complex item (H) figure used a much smaller type size in its labelling, possibly making the item more difficult to read on the screen. As noted in an earlier question, this qualitatively changes the item and may account for the difference in calibration.

Insert Figure 6 about here

Analysis of the standardized differences (Z-scores) between the written and CAT item calibrations, indicates a very strong trend of equivalence. Figure 6 plots item calibrations against the

standardized differences. A Z-score ±2.00 indicates 99% confidence that the item calibrations are different. However, all Z-scores were within ±2.00 and most or all residents were ±1.5 or less. The mean difference of .07 falls well within one standard deviation of the Z-distribution's mean of zero and thus supports this finding of equivalence.

Conclusions

With the increase in popularity of computerized adaptive testing, many organizations are converting written examinations to the CAT format. In this conversion, it is important to be confident that item calibrations are comparable when calculated using data from written and CAT administrations. It is equally important to acknowledge that some items may not function in precisely the same way in written and CAT format because of a change in the examinee's perception or because of some qualitative factor related to the comparability of the item across formats. Our initial investigation found only a few substantial differences in item calibrations across the two administration modes.

Text only MCQ items were found to be equivalently calibrated on written and CAT formats. Similarly, most of the MCQ items with photographic or figure accompaniment in our sample were comparably calibrated. In those few instances where differences were found, the differences could generally be explained by content or other non-statistical, qualitative factors. When images are taken from a printed form and placed onto the screen in a CAT format, careful attention is required to insure that the images are indeed

interpretable. Differences in picture clarity, image size, text complexity may all contribute to the image being qualitatively different. Such substantive issues are extremely important in moving from a written to a CAT format. Where printed photographs and figures are identical to the on screen images, the item calibrations seem as stable across administrative modes as text only MCQ items.

While a number of complex figures were used in the comparisons, further investigation should include the analysis of additional items with complex non-photographic chart or figure accompaniments.

Sources

Bergstrom, B.A. & Lunz, M.E. (In Press). Equivalence of rasch item calibrations and ability estimates across modes of administration. In **Objective Measurement 2**, Ed. M. Wilson. Norbrook, N.J.: Ablex Publishing.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. **Journal of Educational Statistics, (15)** 2, 129-145.

Green, B.F. (1988). Critical problems in computer-based psychological measurement. **Applied Measurement in Education, (3)**, 223-231.

Lord, F.M. (1980). **Applications of Item Response Theory to Practical Testing Problems**. Hillsdale, NJ: Erlbaum.

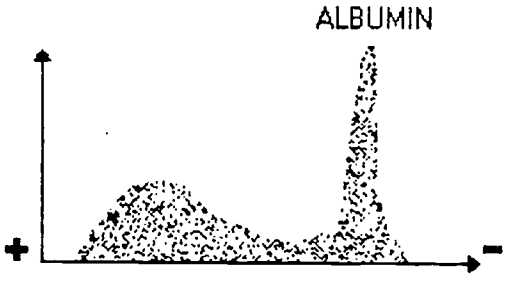
Semejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. **Psychometrika, 42**, 193-198.

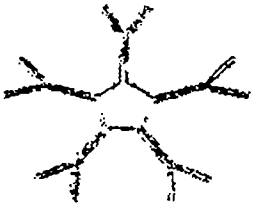
Wright, B.D. and Linacre, J.M. (1991). **Bigsteps** (Computer Program). Chicago: MESA Press.

Wright, B.D. and Stone, M. (1979). **Best Test Design**. Chicago: MESA Press.

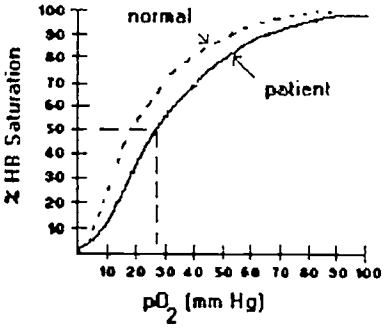
Figure 1: Examples of Figure Complexity Levels

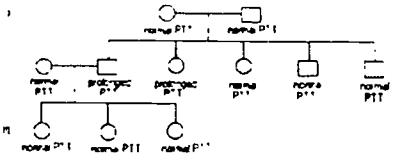
SIMPLE



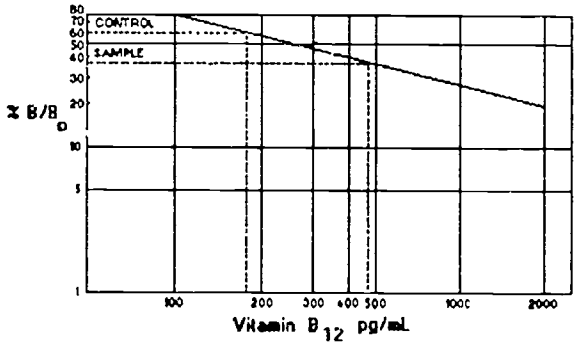


AVERAGE





COMPLEX



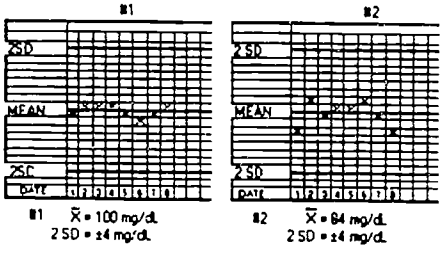


Table 1: Item Calibrations (Text Only MCQ Items)

Test 1 (shown in Figure 2)

Item	--- CAT --- Calib	SE	- Written - Calib	SE	CAT-Written Residual
1	2.57	0.31	2.79	0.03	-0.22
2	2.07	0.23	2.46	0.03	-0.39
3	1.02	0.25	0.35	0.04	0.67
4	0.89	0.27	0.85	0.05	0.04
5	0.45	0.26	0.33	0.03	0.12
6	-0.17	0.22	-0.20	0.04	0.03
7	-0.75	0.43	-1.27	0.06	0.52
8	-0.85	0.27	-0.96	0.05	0.11
9	-1.76	0.31	-1.52	0.05	-0.24
10	-2.19	0.45	-1.55	0.04	-0.64

X = 0.13 SD = 1.56 •

N Items = 10
Range of persons used to calibrate CAT items = 20 - 33
N of persons used to calibrate written items = 1052

Linear Transformation of Written Scores to CAT scale: y=.01 + .763*x*

Table 2: Item Calibrations (Text Only MCQ Items)

Test 2 (shown in figure 3)

Item	--- CAT --- Calib	SE	- Written - Calib	SE	CAT-Written Residual
1	1.59	0.21	1.45	0.03	0.14
2	1.06	0.22	1.00	0.06	0.06
3	1.06	0.29	1.21	0.04	-0.15
4	1.02	0.28	1.01	0.03	0.01
5	0.76	0.32	0.43	0.03	0.33
6	0.51	0.42	1.06	0.04	-0.55
7	0.51	0.33	0.20	0.03	0.31
8	-0.06	0.23	-0.01	0.03	-0.05
9	-0.80	0.28	-0.56	0.04	-0.24
10	-0.84	0.31	-1.00	0.03	0.16

X = 0.48 SD = 0.81 •

N Items = 10
Range of persons used to calibrate CAT items = 20 - 35
N of persons used to calibrate written items = 731

Linear Transformation of Written Scores to CAT scale: y=.25 + .564*x*

• Since written measures were transformed onto the CAT scale, means and standard deviation units for each group are identical.

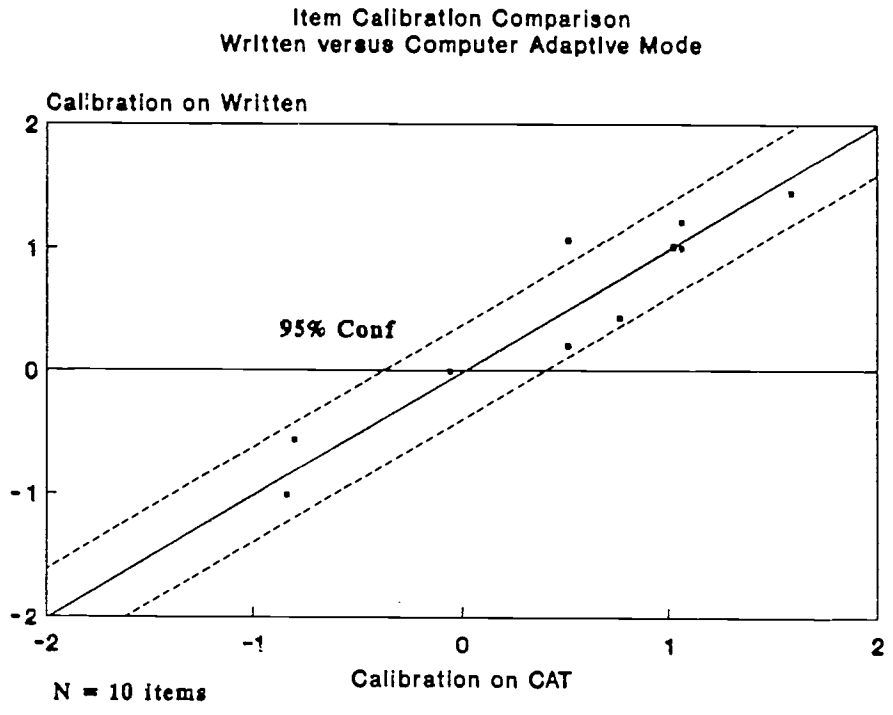


Figure 3

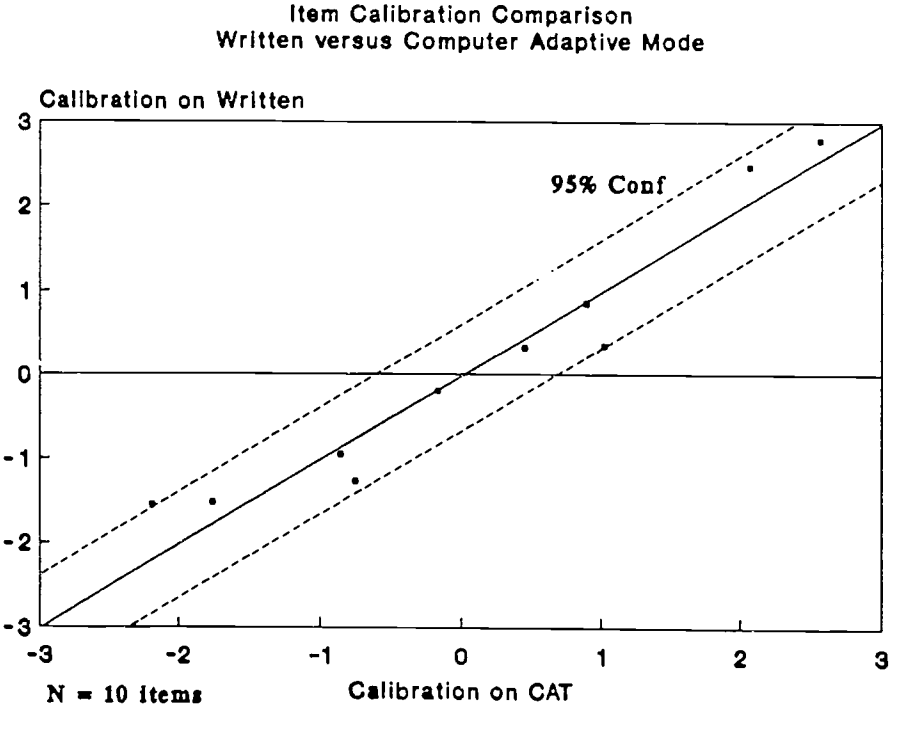


Figure 2

Table 3: Item Calibrations (MCQ w/ Figures or Photographs)

(shown in figure 4)

Item	--- CAT --- Calib	SE	- Written - Calib	SE	CAT-Written Residual	
1a	0.74		0.39	0.05	0.35	
2a	0.64		0.23	0.04	0.41	
3a	-0.38		0.77	0.06	-1.15	†
4a	-0.38		-0.15	0.06	-0.23	
5b	0.10		-0.91	0.06	1.01	†
6b	-0.65		-0.28	0.07	-0.37	
7c	0.07		-0.01	0.05	0.08	
8c	1.46		0.39	0.07	1.07	†
9c	1.02		1.32	0.04	-0.30	
10c	1.17		0.99	0.05	0.18	
11c	0.88		1.12	0.05	-0.24	
12c	1.08		0.91	0.06	0.17	
13c	0.37		1.28	0.07	-0.91	†

N Items:

Test a = 4
Test b = 2
Test c = 7

Range of persons used to calibrate CAT items = 17 - 29

N of persons used to calibrate written items:

Test a = 1052
Test b = 731
Test c = 641

Linear Transformation of Written Scores to CAT scale:

Test a: $y=.01 + .763x$ [X = 0.13 SD = 1.56 •]
Test b: $y=.25 + .564x$ [X = 0.48 SD = 0.81 •]
Test c: $y=.02 + .827x$ [X = 1.05 SD = 1.17 •]

† See four items outside the 95% confidence band (Figure 4)

• Since written measures were transformed onto the CAT scale, means and standard deviation units for each group are identical.

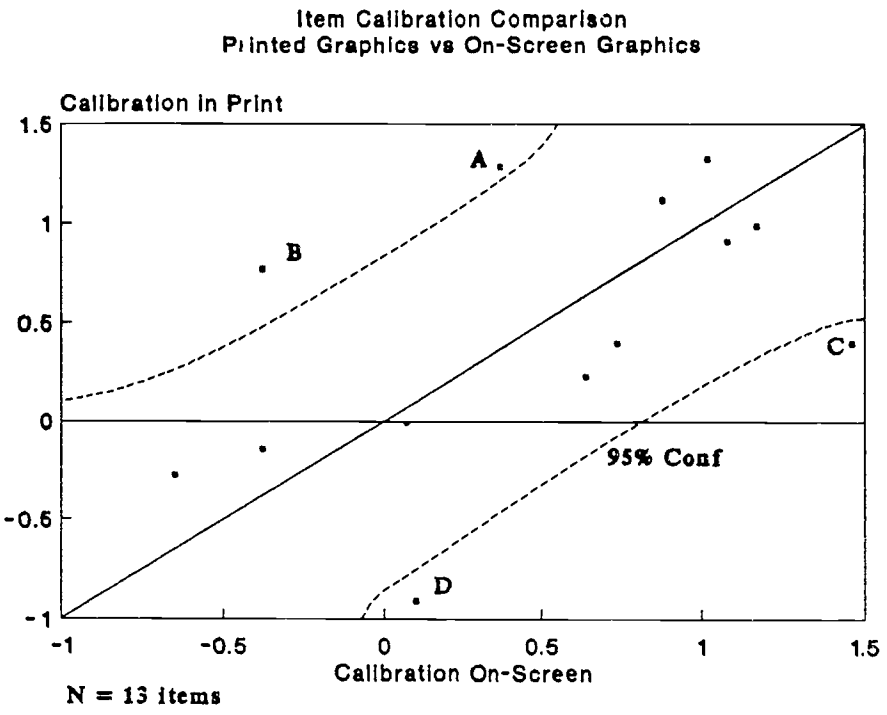

Item Calibration Comparison
Printed Graphics vs On-Screen Graphics
Calibration in Print
1.5
1
0.6
0
-0.6
-1
A
B
C
D
95% Conf
-1
-0.5
0
0.5
1
1.5
Calibration On-Screen
N = 13 items

Figure 4

Table 4: Item Calibrations (MCQ w/ Figures)

(shown in figure 5)

Item	--- CAT --- Calib	SE	- Written - Calib	SE	CAT-Written Residual
1a	-0.16	0.44	0.19	0.06	-0.35
2a	-0.12	0.46	-0.57	0.05	0.45
3a	0.27	0.78	-1.44	0.10	1.71 †
4a	-0.36	0.31	-0.12	0.07	-0.24
5a	0.19	0.21	0.13	0.04	-0.06
6a	2.07	0.75	1.12	0.05	-0.95 †
7a	0.20	0.45	-0.22	0.06	0.42
8b	-0.83	0.43	-0.20	0.03	-0.63
9b	0.01	0.20	-0.09	0.05	0.10
10b	-0.92	0.21	-0.96	0.04	0.04
11b	0.24	0.37	0.84	0.06	-0.60
12b	0.17	0.42	-0.49	0.07	0.66
13b	0.65	0.32	0.31	0.05	0.34
14b	0 29	0.39	0.77	0.05	-0.48
15b	0.14	0.41	-0.26	0.06	0.40
16c	0.75	0.34	0.86	0.04	-0.11
17c	0.15	0.50	-0.20	0.05	0.35
18c	-0.09	0.52	0.25	0.07	-0.34
19c	-0.89	0.55	-0.27	0.04	-0.62
20c	0.75	0.38	0.47	0.04	0.28
21c	0.12	0.48	0.75	0.05	-0.63

N Items:

Test a = 7
Test b = 8
Test c = 6

Range of persons used to calibrate CAT items = 17 - 31

N of persons used to calibrate written items:

Test a = 1052
Test b = 731
Test c = 641

Linear Transformation of Written Scores to CAT scale:

Test a: $y=.01 + .763x$ [X = 0.13 SD = 1.56 *]
Test b: $y=.25 + .564x$ [X = 0.48 SD = 0.81 *]
Test c: $y=.02 + .827x$ [X = 1.05 SD = 1.17 *]

† See two items outside the 95% confidence band (Figure 5)

* Since written measures were transformed onto the CAT scale, means and standard deviation units for each group are identical.

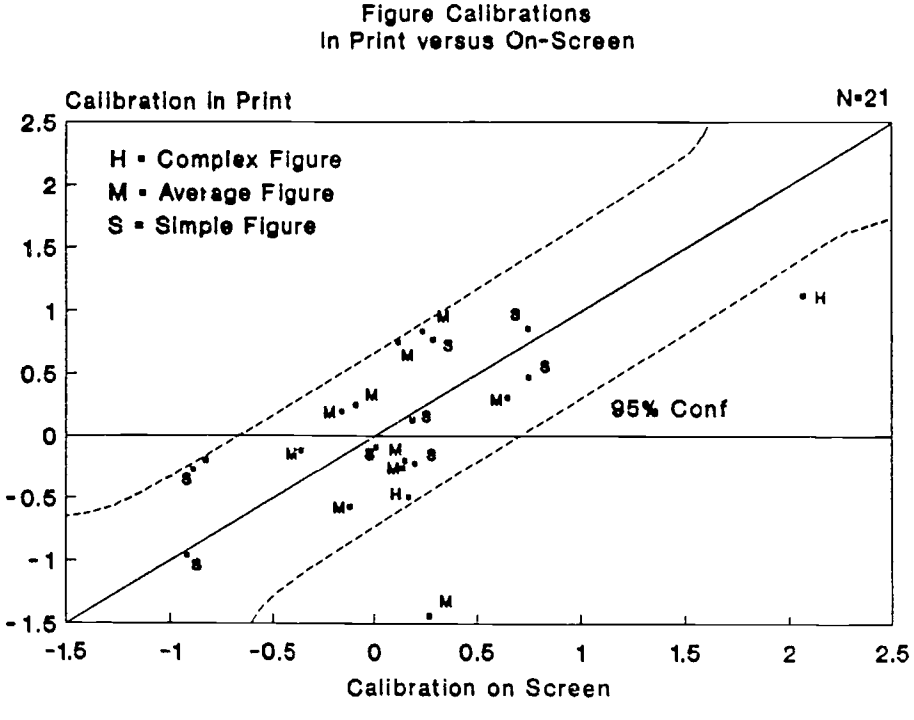


Figure 5

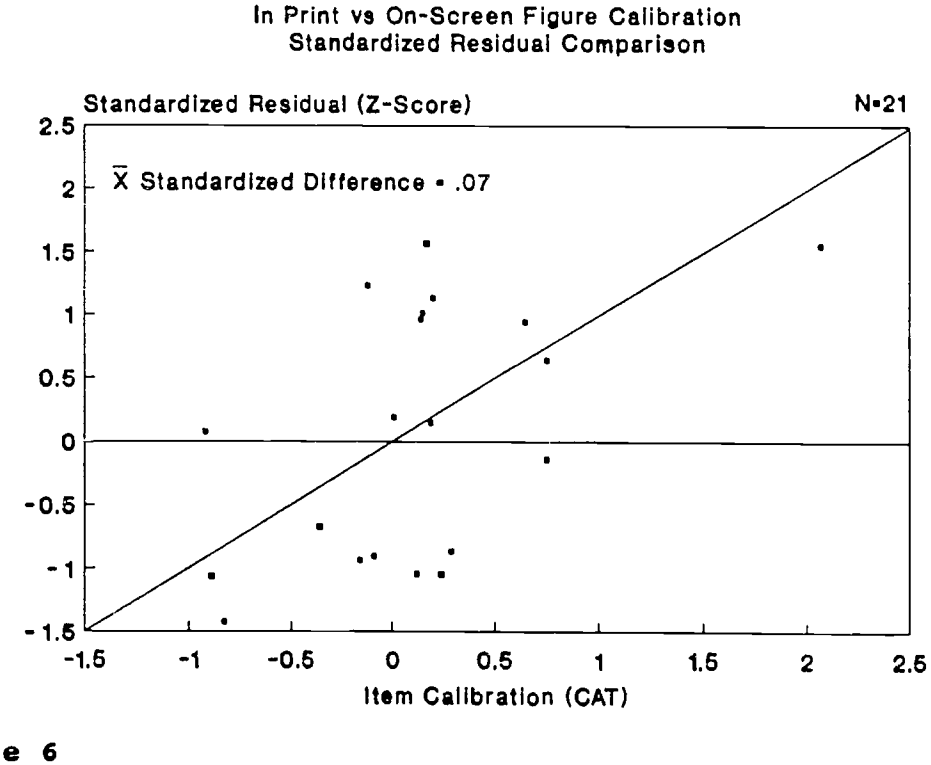


Figure 6